Statistical Modeling of Social Networks

Mark S. Handcock

Department of Statistics
University of California - Los Angeles

Joint work with

Ian M. Fellows David Hunter
Krista Gile Pavel Krivitsky

Supported by NIH NIDA Grant DA012831, NICHD Grant HD041877, NSF award MMS-0851555 and the DoD ONR MURI award N00014-08-1-1015.

Working Papers available at
`http://www.stat.ucla.edu/~handcock`
`http://statnet.org`

SAMSI Workshop in Computational Advertising, August 6-8 2012

Relevance of the statistical analysis of networks

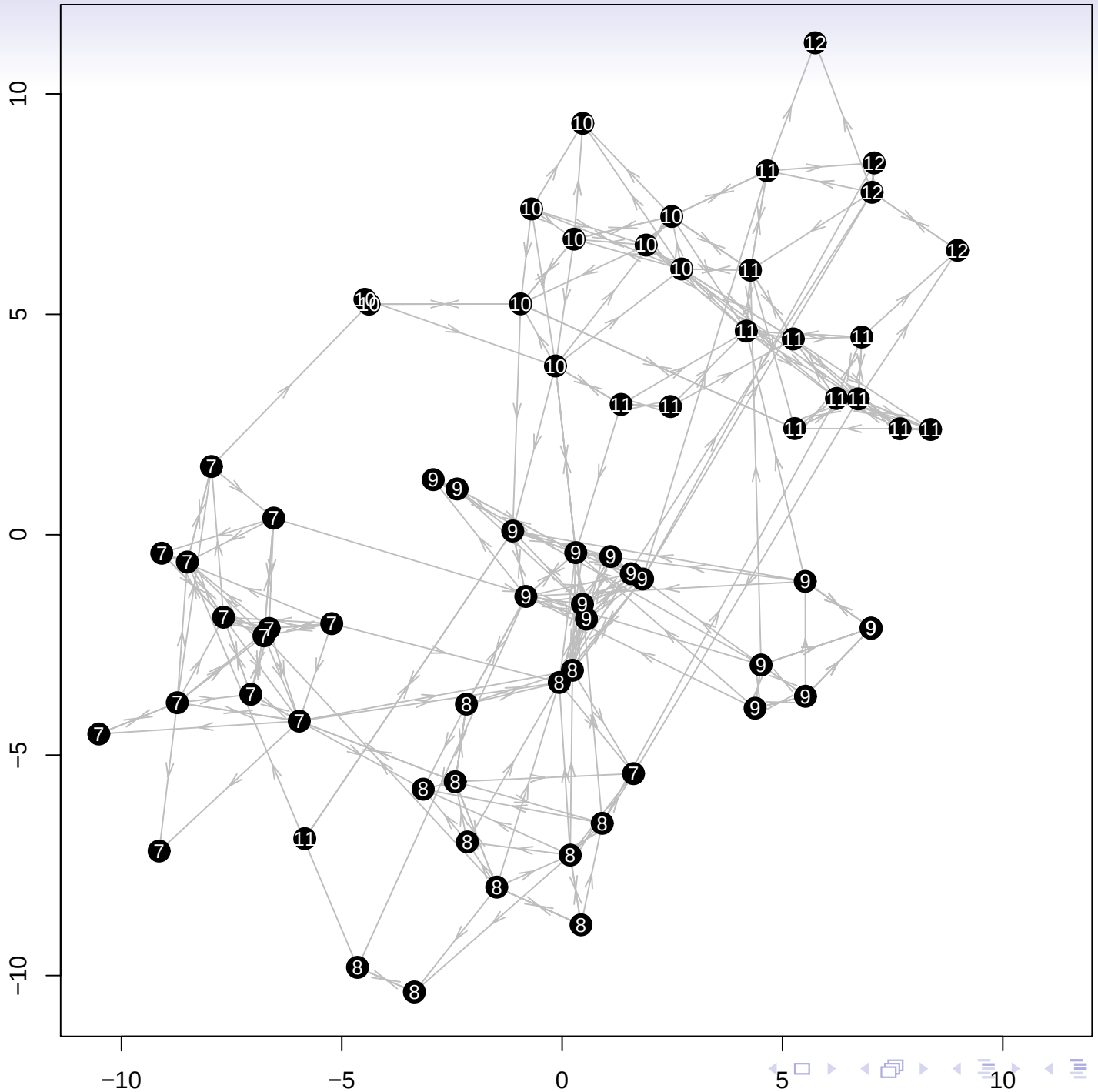
- Understanding the structure of social relations has been the focus of the social sciences
  - *social structure:* a system of social relations tying distinct social entities to one another
  - Understanding social structure is important to many aspects of computational advertising.
  - e.g. Work starting in 1940s at the Columbia Bureau of Applied Social Research, lead by Paul Lazarsfeld.
- Networks are widely used to represent data on relations between interacting actors or nodes.
- Attempt to represent the structure in social relations via networks
- The study of social networks is multi-disciplinary
  - plethora of terminologies
  - varied objectives, multitude of frameworks

Deep literatures available

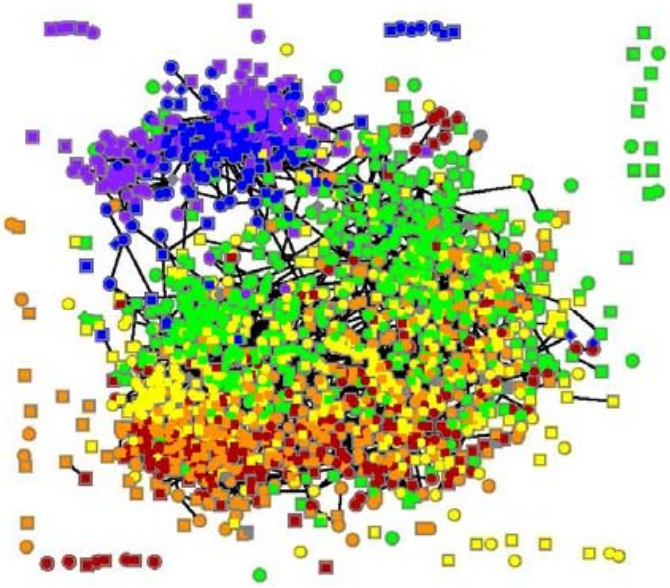
- Social networks community (Heider 1946; Frank 1972; Holland and Leinhardt 1981)
- Statistical Networks Community (Frank and Strauss 1986; Snijders 1997, Fienberg)
- Spatial Statistics Community (Besag 1974)
- Statistical Exponential Family Theory (Barndorff-Nielsen 1978)
- Graphical Modeling Community (Lauritzen and Spiegelhalter 1988, . . .)
- Machine Learning Community (Jordan, Jensen, Xing, ...)
- Physics and Applied Math (Newman, Watts, . . .)

Examples of Friendship Relationships

- The National Longitudinal Study of Adolescent Health
  $\Rightarrow$ `www.cpc.unc.edu/projects/addhealth`
  - "Add Health" is a school-based study of the health-related behaviors of adolescents in grades 7 to 12.
- Each nominated up to 5 boys and 5 girls as their friends
- 160 schools: Smallest has 69 adolescents in grades 7–12


10
5
0
−5
−10
−10
−5
0
5
10

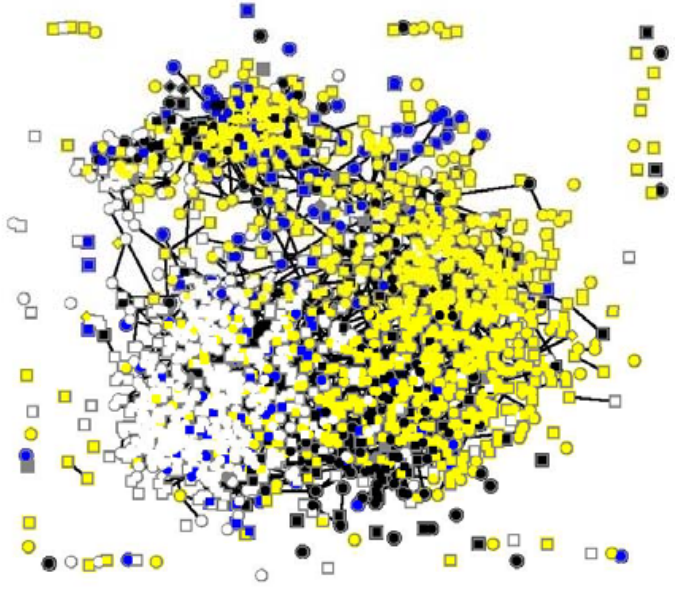
School Community Stratum 44
mutual friendships by Grade



2209 Students

- **Grade 7**
- **Grade 8**
- **Grade 9**
- **Grade 10**
- **Grade 11**

School Community Stratum 44
mutual friendships by Race



2209 Students

- **White** (non-Hispanic)
- **Black** (non-Hispanic)
- **Hispanic** (of any race)
- **Asian / Native Am / Other** (non-Hispanic)
- **Race NA**

Features of Many Social Networks

- *Mutuality* of ties
- *Individual heterogeneity* in the propensity to form ties
- *Homophily* by actor attributes

  $\Rightarrow$ Lazarsfeld and Merton, 1954; Freeman, 1996; McPherson et al., 2001
  - higher propensity to form ties between actors with similar attributes
    e.g., age, gender, geography, major, social-economic status
  - attributes may be observed or unobserved
- *Transitivity* of relationships
  - friends of friends have a higher propensity to be friends
- *Balance* of relationships $\Rightarrow$ Heider (1946)
  - people feel comfortable if they agree with others whom they like
- *Context* is important $\Rightarrow$ Simmel (1908)
  - triad, not the dyad, is the fundamental social unit

Perspectives to keep in mind

- Network-specific versus Population-process
  - *Network-specific*: interest focuses only on the actual network under study
  - *Population-process*: the network is part of a population of networks and the latter is the focus of interest
    - the network is conceptualized as a realization of a social process
- The choice of models depends on the objectives
  - The complexity of most network processes precludes complete modeling
  - We choose those aspects of the network we represent and model them well

Statistical Models for Social Networks

Notation

A *social network* is defined as a set of n social "actors", a social relationship between each pair of actors, and a set of variables on those actors/pairs.

$$Y_{ij} = \begin{cases} 1 & \text{relationship from actor } i \text{ to actor } j \\ 0 & \text{otherwise} \end{cases}$$

- call $Y \equiv [Y_{ij}]_{n \times n}$ a *sociomatrix*
  - a $N = n(n-1)$ binary array
- X be $n \times q$ matrix of actor characteristics
- The basic problem of stochastic modeling is to specify a distribution for X, Y i.e., $P(Y = y, X = x)$

Rich Models for Networks

Let $\mathcal{Y}$ be the sample space of Y e.g. $\{0,1\}^N$
and $\mathcal{X}$ be the sample space of X.
Use exponential-family models for the multivariate distribution of Y:

$$P_\eta(Y = y|X = x) = \frac{\exp\{\eta \cdot g(y|x)\}}{\kappa(\eta, x, \mathcal{Y})} \qquad y \in \mathcal{Y}, \quad x \in \mathcal{X}$$

Frank and Strauss (1986)

- $\eta \in \Lambda \subset R^q$ q-vector of parameters
- $g(y|x)$ q-vector of *network statistics*.
  $\Rightarrow$ $g(Y|x)$ are jointly sufficient for the model
- For a "saturated" model-class $q = |\mathcal{Y}| - 1$ e.g. $2^N - 1$
- $\kappa(\eta, x, \mathcal{Y})$ distribution normalizing constant

$$\kappa(\eta, x, \mathcal{Y}) = \sum_{y \in \mathcal{Y}} \exp\{\eta \cdot g(y|x)\}$$

Simple model-classes for social networks

Homogeneous Bernoulli graph (Erdős-Rényi model)

- Y_{ij} are independent and equally likely
  with log-odds $\eta = \text{logit}[P_\eta(Y_{ij} = 1)]$

$$P_\eta(Y = y) = \frac{e^{\eta \sum_{i,j} y_{ij}}}{\kappa(\eta, \mathcal{Y})} \qquad y \in \mathcal{Y}$$

  where $q = 1$, $g(y) = \sum_{i,j} y_{ij}$, $\kappa(\eta, \mathcal{Y}) = [1 + \exp(\eta)]^N$
  (There is no x here).
- homogeneity means it is unlikely to be proposed as a model for real phenomena

Dyad-independence models with attributes

- Y_{ij} are independent but depend on dyadic covariates $x_{k,ij}$

$$P_\eta(Y = y|X = x) = \frac{e^{\sum_{k=1}^{q} \eta_k g_k(y|x)}}{\kappa(\eta, x, \mathcal{Y})} \qquad y \in \mathcal{Y}$$

$$g_k(y|x) = \sum_{i,j} x_{k,ij} y_{ij}, \qquad k = 1, \ldots, q$$

$$\kappa(\eta, x, \mathcal{Y}) = \prod_{i,j} [1 + \exp(\sum_{k=1}^{q} \eta_k x_{k,ij})]$$

Of course,

$$logit[P_\eta(Y_{ij} = 1)] = \sum_k \eta_k x_{k,ij}$$

Generative Theory for Network Structure

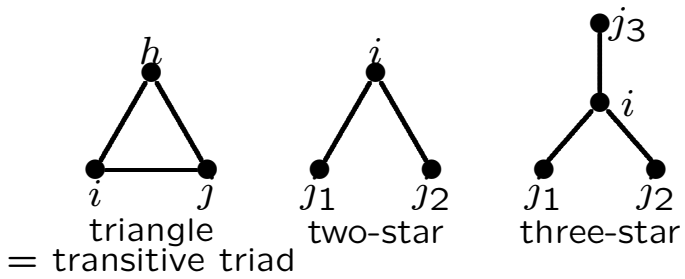
Actor Markov statistics

$\Rightarrow$ Frank and Strauss (1986)

– motivated by notions of "symmetry" and "homogeneity"
conditionally independent given the rest of the network

$\Rightarrow$ analogous to nearest neighbor ideas in spatial modeling

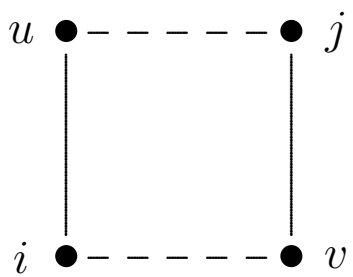
- Degree distribution: $\mathrm{d}_k(y) =$ proportion of actors of degree k in y.
- triangles: $\mathrm{triangle}(y) =$
number of triads that form a complete sub-graph in y.



More General mechanisms motivated by conditional independence

$\Rightarrow$ Pattison and Robins (2002), Butts (2005)
$\Rightarrow$ Snijders, Pattison, Robins and Handcock (2006)

- Y_{uj} and Y_{iv} in Y are conditionally independent given the rest of the network if they could not produce a cycle in the network



Partial conditional dependence when four-cycle is created

This produces features on configurations of the form:

- edgewise shared partner distribution: $\mathrm{esp}_k(y) =$ proportion of edges between actors with exactly k shared partners $k = 0, 1, \ldots$

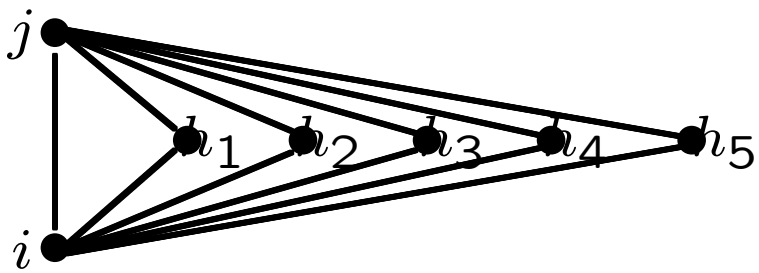


Figure: The actors in the non-directed (i, j) edge have 5 shared partners

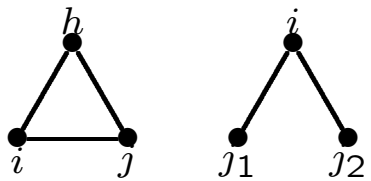
- dyadwise shared partner distribution: $\mathrm{dsp}_k(y) =$ proportion of dyads with exactly k shared partners $k = 0, 1, \ldots$

Structural Signatures

– identify social constructs or features
– based on intuitive notions or partial appeal to substantive theory

- Clusters of edges are often *transitive*:
  Recall $\text{triangle}(y)$ is the number of triangles amongst triads

$$\text{triangle}(y) = \frac{1}{\binom{g}{3}} \sum_{\{i,j,k\} \in \binom{g}{3}} y_{ij} y_{ik} y_{jk}$$



A closely related quantity is the
proportion of triangles amongst two-stars

$$C(y) = \frac{3 \times \text{triangle}(y)}{\text{two-star}(y)}$$

mean clustering coefficient

Extensive development of conditional models

- Classes of $g(y|x)$ (Generative Theory, Structural signatures)
- Inference on the loglikelihood function,

$$\ell(\eta|y_{\text{obs}}; x_{\text{obs}}) = \eta \cdot g(y_{\text{obs}}|x_{\text{obs}}) - \log \kappa(\eta|x_{\text{obs}})$$

$$\kappa(\eta|x_{\text{obs}}) = \sum_{z \ in \mathcal{Y}} \exp\{\eta \cdot g(z|x_{\text{obs}})\}$$

- For computational reasons, approximate the likelihood via Markov Chain Monte Carlo (MCMC)

Exponential-family Random Network Models

Joint modeling of Y and X

Let $\mathcal{N}$ be the sample space of Y, X

Model the multivariate distribution of Y, X via the form:

$$P_\eta(Y = y, X = x) = \frac{\exp\{\eta \cdot g(y, x)\}}{\kappa(\eta, \mathcal{N})} \qquad y, \quad x \in \mathcal{N}$$

- $\eta \in \Lambda \subset R^q$ q-vector of parameters
- $g(y, x)$ q-vector of *network statistics*.
  $\Rightarrow$ $g(Y, X)$ are jointly sufficient for the model
- $\kappa(\eta, \mathcal{N})$ distribution normalizing constant

$$\kappa(\eta, \mathcal{N}) = \int_{y,\ x \in \mathcal{N}} \exp\{\eta \cdot g(y, x)\} \cdot dP_0(y, x)$$

where $P_0(\cdot, \cdot)$ is a reference measure.

Interesting model-classes of ERNM

Relationship to ERGM and Random Fields

Let $\mathcal{N}(x) = \{y : (x, y) \in \mathcal{N}\}$ and $\mathcal{N}(y) = \{x : (x, y) \in \mathcal{N}\}$

$$\text{ERGM} \qquad P(Y = y|X = x; \eta) = \frac{1}{c(\eta; x)} e^{\eta \cdot g(x,y)} \quad y \in \mathcal{N}(x)$$

$$\text{Gibbs measure} \qquad P(X = x|Y = y; \eta) = \frac{1}{c(\eta; y)} e^{\eta \cdot g(x,y)} \quad x \in \mathcal{N}(y)$$

- The first model is the ERGM for the network conditional on the nodal attributes.
- The second model is an exponential-family for the field of nodal attributes conditional on the network

Conditional Models

The model can be expressed as

$$P(X = x, Y = y|\eta) = P(Y = y|X = x|\eta)P(X = x|\eta)$$

where

$$\begin{aligned} P(Y = y|X = x;\eta) &= \frac{1}{c(\eta;x)} e^{\eta \cdot g(x,y)} \quad y \in \mathcal{N}(x) \\ P(X = x|\eta) &= \frac{c(\eta;x)}{c(\eta)} \quad x \in \mathcal{X} \end{aligned}$$

- The first sub-model is the ERGM for the network conditional on the nodal attributes.
- The second sub-model is the marginal representation of the nodal attributes and is not necessarily an exponential-family with canonical parameter η.
- This decomposition makes it clear why the conditional modeling of Y given X via ERGM differs from the joint modeling of Y and X via ERNM.

Latent variable models for social structure

- less parametric, but more flexible than ERGM-type models
- model an underlying latent "social space" of actors
  - Latent space models: Hoff, Raftery and Handcock (2002)
    Hoff (2003, 2004 ,...)
  - Latent class and space models: Tantrum, Handcock, Raftery (2004)
  - GLM and actor heterogeneity: Krivitsky et al (2009)
- Hierarchical model for the network:
  - Actors i and j are an unknown distance apart in social space
  - Conditional on the distances the ties are independent

  Let:
  - $\{\delta_i\}$ individual propensity of the actors to form ties
  - $\{\gamma_i\}$ individual propensity of the actors to receive ties
  - $\{z_i\}$ be the positions of the actors in the social space $\mathbf{R}^k$
  - $\{x_{i,j}\}$ denote observed characteristics that may be dyad-specific and vector-valued

  Specifically:

$$\text{log odds}(Y_{ij} = 1 | z_i, z_j, x_{ij}, \beta) = \beta^T x_{ij} - |z_i - z_j| + \delta_i + \gamma_j$$

  where β denotes parameters to be estimated.

Model-based Clustering of Social Networks

- Model the latent positions as clustered into G groups:

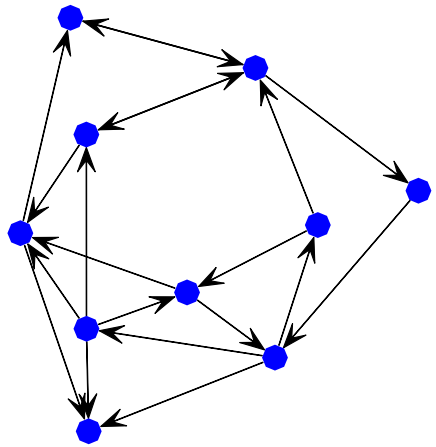
$$z_i \overset{\text{i.i.d.}}{\sim} \sum_{g=1}^{G} \lambda_g \text{MVN}_d(\mu_g, \sigma_g^2 I_d)$$

  - Spherical covariance motivated by invariance
- captures position, transitivity, homophily on attributes, and clustering
- captures individual propensities to form and receive ties

$$\begin{aligned} \delta_i &\overset{\text{i.i.d.}}{\sim} \text{N}(0, \sigma_\delta^2) \quad i = 1, \ldots, n, \\ \gamma_i &\overset{\text{i.i.d.}}{\sim} \text{N}(0, \sigma_\gamma^2) \quad i = 1, \ldots, n, \end{aligned}$$

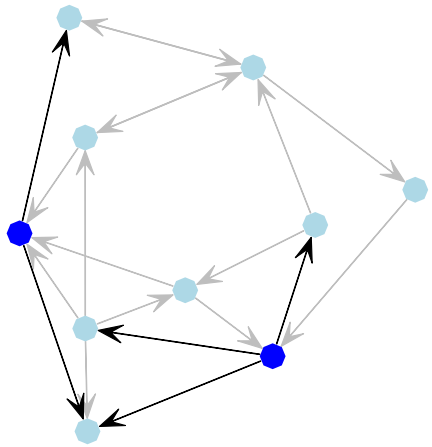
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



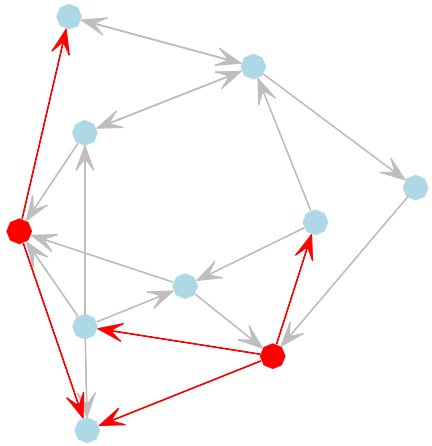
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



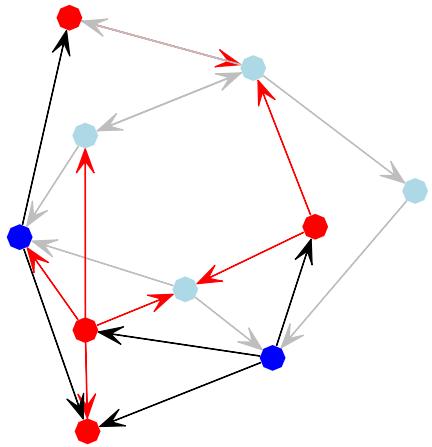
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



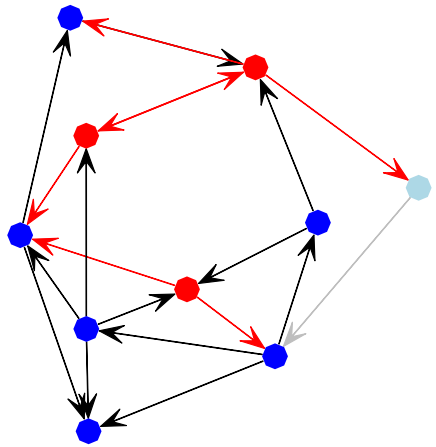
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



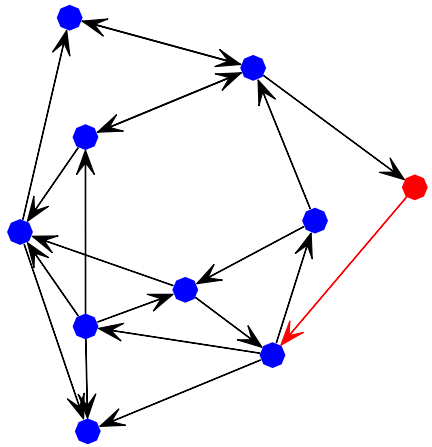
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



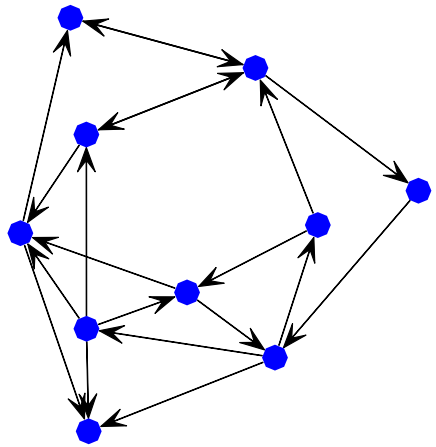
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



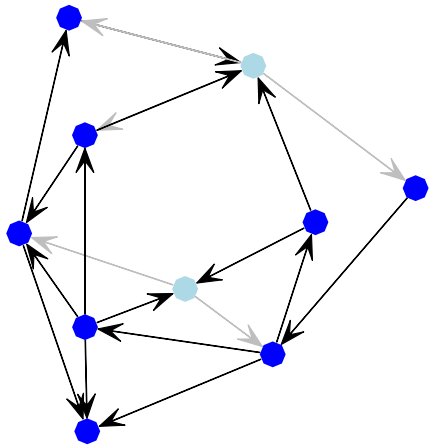
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



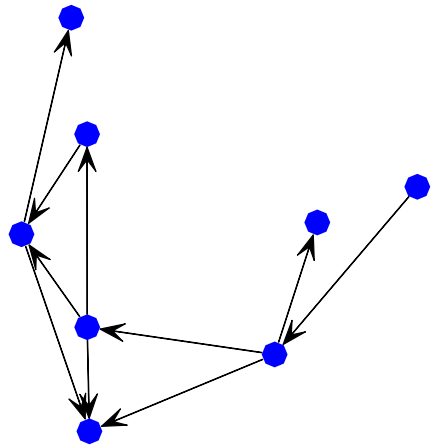
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



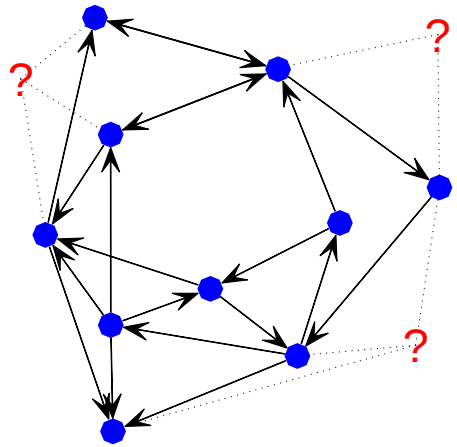
Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



Partial Observation of Social Networks

- **Sampling Design:** Choose which part to observe:
  "Ask 10% of employees about their collaborations"
  - Egocentric
  - Adaptive
- **Out-of-design Missing Data:**
  "Try to survey the whole company, but someone is out sick"
- **Boundary Specification Problem:**
  "Should a contractor be considered a part of the company?"



Design-based Inference for Describing Structure

- Approach:
  - Make probability statements about the relations in the full network based on the observed part of the network
  - Weight each observation by the inverse of probability of being sampled
- Advantages:
  - Requires no assumptions about network structure
- Disadvantages:
  - Requires full knowledge of sampling mechanism, and sampling probabilities
  - Difficult to conduct complex analysis

Observable sampling probabilities under various sampling schemes

Sampling Scheme	Nodal Probabilities π_i		Dyadic Probabilities π_{ij}	
	Undirected	Directed	Undirected	Directed
Ego-centric	X	X	X	X
One-Wave	X			
$k-$Wave, $1 < k < \infty$				
Saturated	X			

"X" indicates observable sampling probabilities

Social Network Modeling for Understanding Processes

- Approach:
  - Make probability statements about the social forces that could account for the network
  - Create complex super-population model for the relational information
- Advantages:
  - Flexible Models to answer complex questions
- Disadvantages:
  - Assumes chosen model form is accurate
  - Computationally expensive for complex models
  - Assumptions about the sampling/missing process
  - Initially, only fit to fully observed networks

Fitting Models to Networks with Incomplete Data

- Two types of units: nodes and relational structures
- Sampling typically on nodes, inference on relational structures

- Extend and adapt methods from survey sampling and missing data literature (Thompson and Seber, 1996, Little and Rubin, 2002)
- Extend former work on partially-observed network data (Frank, 1971, Frank and Snijders, 1994, Thompson and Frank, 2000)
- Novel Methods: Full range of stochastic models; expand model-checking (Handcock and Gile, 2007, Gile 2008, 2009, Gile and Handcock, 2006, 2009, Fellows and Handcock 2011)

- Key Point: require that statistical properties of unobserved relations do not depend on unobserved characteristics, given what was observed

Fitting Models to Partially Observed Social Network Data

- Two types of data: Observed relations (y_{obs}), and indicators of units sampled (D).

$$
\begin{aligned}
L(\boldsymbol{\eta}, \psi) &\equiv P(Y_{obs} = y_{obs}, D|\boldsymbol{\eta}, \psi) \\
&= \sum_{y_{unobs}} P(Y_{obs} = y_{obs}, Y_{unobs} = y_{unobs}, D|\boldsymbol{\eta}, \psi) \\
&= \sum_{y_{unobs}} P(D|Y_{obs} = y_{obs}, Y_{unobs} = y_{unobs}, \psi) P_{\boldsymbol{\eta}}(Y_{obs} = y_{obs}, Y_{unobs} = y_{unobs})
\end{aligned}
$$

- $\boldsymbol{\eta}$ is the model parameter
- ψ is the sampling parameter

When can we "ignore" the sampling process?

Adaptive Sampling Designs

- We call a sampling design *adaptive* if:

$$P(D = d|Y_{obs}, Y_{mis}, \psi) = P(D = d|Y_{obs}, \psi) \qquad \forall y \in \mathcal{Y}.$$

that is, it uses information collected during the survey to direct subsequent sampling, but the sampling design depends only on the observed data.

- adaptive sampling designs satisfy a condition called "*missing at random*" by Rubin (1976) in the context of missing data.

- **Result:** standard network sampling designs such as conventional, single wave and multi-wave link-tracing sampling designs are adaptive
  $\Rightarrow$ Thompson and Frank (2000), Handcock and Gile (2007).

When is sampling adaptive?

Examples of adaptive sampling:

- Individual sample, sample based on observed things like race, sex, and age that we know.
- Link-tracing sample starting with a adaptive sample with follow-up based on observed relations with others in the sample, as well as things like race and sex and age.
- Link-tracing with probability proportional to number of partners is adaptive!

Examples of non-adaptive (not missing at random) sampling:

- Individual sample based on unobserved properties of non-respondents - like infection status or illicit activity.
- Link-tracing sample starting where links are followed dependent on unobserved properties of alters.

Adaptive Sampling Designs and their Amenable Models

Definition: Consider a sampling design governed by parameter $\psi \in \Psi$ and a stochastic network model $P_{\boldsymbol{\eta}}(Y = y)$ governed by parameter $\boldsymbol{\eta} \in \Xi$. We call the sampling design *amenable to the model* if the sampling design is adaptive and the parameters ψ and $\boldsymbol{\eta}$ are distinct.

Result: If the sampling design is amenable to the model the likelihood for $\boldsymbol{\eta}$ and ψ is

$$L[\boldsymbol{\eta}, \psi | Y_{obs} = y_{obs}, D = d] \propto L[\psi | D = d, Y_{obs} = y_{obs}] L[\boldsymbol{\eta} | Y_{obs} = y_{obs}]$$

Adaptive Sampling Designs and their Amenable Models

Definition: Consider a sampling design governed by parameter $\psi \in \Psi$ and a stochastic network model $P_{\boldsymbol{\eta}}(Y = y)$ governed by parameter $\boldsymbol{\eta} \in \Xi$. We call the sampling design *amenable to the model* if the sampling design is adaptive and the parameters ψ and $\boldsymbol{\eta}$ are distinct.

Result: If the sampling design is amenable to the model the likelihood for $\boldsymbol{\eta}$ and ψ is

$$L[\boldsymbol{\eta}, \psi | Y_{obs} = y_{obs}, D = d] \propto L[\psi | D = d, Y_{obs} = y_{obs}] L[\boldsymbol{\eta} | Y_{obs} = y_{obs}]$$

sampling design likelihood $\times$ *face-value likelihood*

$$L[\psi | D = d, Y_{obs} = y_{obs}] = P(D | Y_{obs} = y_{obs}, \psi)$$

$$L[\boldsymbol{\eta} | Y_{obs} = y_{obs}] = \sum_{y_{unobs}} P_{\boldsymbol{\eta}}(Y_{obs} = y_{obs}, Y_{unobs} = y_{unobs})$$

Adaptive Sampling Designs and their Amenable Models

Result: If the sampling design is *not* amenable to the model the likelihood for $\boldsymbol{\eta}$ and ψ is

$$L(\boldsymbol{\eta}, \psi) = \sum_{y_{unobs}} P(D|Y_{obs} = y_{obs}, Y_{unobs} = y_{unobs}, \psi) P_{\boldsymbol{\eta}}(Y_{obs} = y_{obs}, Y_{unobs} = y_{unobs})$$

and the design will need to be represented.

Clearly $P(D|Y, \psi)$ can be modeled when it is unknown.

Conclusions and Challenges

- Network models are a very constructive way to represent (social) theory (e.g., homophily, transitivity, heterogeneity)
- The models can be used to compare the predictions of social theory
- Simple models are being used to capture structural properties
- The choice of models depends on the objectives
  - The complexity of most network processes precludes complete modeling
  - We choose those aspects of the network we represent and model them well
- The inclusion of attributes is very important
  - actor attributes
  - dyad attributes e.g. homophily, race, location
  - structural terms e.g. transitive homophily